Using VARI Model to Forecast Climate Phenomena in Big Data Era

Ajeng Berliana Salsabila, Sarah Sutisna and Sri Purwani

Abstract—**Climate change is currently occurring, with significant impacts on various aspects of human life. To understand and predict the phenomena, the use of statistical model is essential to describe the relationships between various climate variables. A commonly used model is the Vector Autoregressive (VAR) which enables the simultaneous analysis of the dependence and interaction between several climate variables. It is considered a combination of autoregressive (AR) model of the same order. Therefore, this research aimed to forecast climate phenomena in West Java, focusing on the rainfall variable using the VAR model. Rainfall data was obtained from NASA POWER, representing a large dataset categorized as big data. Data Analytics Lifecycle method was used to address the challenges associated with big data. Ordinary Least Squares (OLS) were used as the parameter estimation method within the VAR model. The results obtained using R software showed that the rainfall data had a non-stationary pattern, requiring a differencing process such as VARI model. The Mean Absolute Percentage Error (MAPE) value was less than 20%, showing that forecasting climate phenomena using VARI model was within the accurate category.**

Index Terms—**vector autoregressive integrated (VARI), climate, big data, data analytics lifecycle.**

I. Introduction

CLIMATE is the long-term patterns of temperature, precipitation, humidity, wind, and other atmospheric conditions in a specific region or across the planet [1]. It plays a critical role in determining the natural environment and human societies. Moreover, climate changes can significantly impact ecosystems, including variations in species distribution, disruptions to food webs, and extreme weather events such as droughts, floods, and hurricanes [2], affecting corporate bonds [3]. Climate change caused by human activities such as burning fossil fuels and deforestation, is a significant global challenge affecting economies, human health, and the planet [4]. Meteorology, Climatology, and Geophysical Agency (BMKG) [5] stated that 2016 was the hottest year for Indonesia, with an anomaly value of 0.8°C throughout the observation period from 1981 to 2020. Meanwhile, 2020 was the second hottest year with an anomaly value of 0.7°C, and 2019 ranked third at 0.6°C. The World Meteorological Organization (WMO) [6] has also released global average temperature information for comparison. The latest WMO report on climate published in early 2023 stated that 2022 was ranked 6*th* hottest year in the world, while 2015-2022 was 8*th*.

Climate data can be represented as a time series, a sequence of data points recorded at regular intervals over time. Variables such as temperature, rainfall, and atmospheric pressure can be measured regularly and analyzed as time series data to identify trends and patterns over time. Time series data can also be used to develop model, which simulates the Earth's climate system and forecast conditions based on different scenarios of greenhouse gas emissions and other factors. As a time series, these data are essential for understanding past climate conditions, monitoring current situations, and predicting future change impacts on the environment and human societies. When climate data is continuously sorted occasionally, it becomes big data which are vast and complex datasets. These data are beyond the ability of traditional processing and analysis tools to manage and analyze effectively. The term "big" is relative and can vary depending on the context [7]. Generally, big data refers to datasets that are extremely large, diverse, and rapidly changing to be managed and analyzed using traditional methods. Big data can come from various sources, such as social media, scientific research, financial transactions, and weblogs. It is typically characterized by volume, velocity, and variety, known as the three Vs [8]. The analysis often requires specialized tools and methods, such as distributed computing, machine learning, and natural language processing to extract insights and value from data [9].

Time series is a sequence of data points recorded at regular intervals to identify patterns, trends, and relationships between different variables. Based on stationarity, data is divided into stationary and non-stationary time series [10]. According to the number of variables, it is categorized into univariate and multivariate time series as observations collected at equally spaced intervals of time with only a single variable. One example of stationary univariate time series model is Autoregressive (AR). Meanwhile, non-stationary is Autoregressive Integrated (ARI), which refers to a set of observations collected over time with two or more

Manuscript received July 5, 2023; revised September 17, 2024.

This work was supported in part by the Rector and the Directorate of Research and Community Service (DRPM) Universitas Padjadjaran.

Ajeng Berliana Salsabila is a postgraduate student in the Mathematics Doctoral Program, Faculty of Mathematics and Natural Sciences, Universitas Padjadjaran, Sumedang 45363, Indonesia (e-mail: ajeng18004@mail.unpad.ac.id).

Sarah Sutisna is a graduate of Magister of Mathematics, Faculty of Mathematics and Natural Sciences, Universitas Padjadjaran, Sumedang 45363, Indonesia (e-mail: sarah19005@mail.unpad.ac.id).

Sri Purwani is a lecturer in the Mathematics Department, Faculty of Mathematics and Natural Sciences, Universitas Padjadjaran, Sumedang 45363, Indonesia (corresponding author; e-mail: sri.purwani@unpad.ac.id).

interrelated variables. In dataset, variables are considered to be dependent or independent. An example of stationary multivariate time series model is Vector Autoregressive (VAR), while Vector Autoregressive Integrated (VARI) is non-stationary.

According to Wei [10], each variable in VAR model is a function of the past values and the other variables' past values. Model assumes that each variable in the system affects one another without causal relationship. This shows that model allows for the possibility of feedback loops and interactions between variables. One of the advantages of VAR model is the ability to capture the dynamic relationships between variables. This can be useful for forecasting future values of the system and understanding underlying drivers. It can also be estimated using maximum possibility estimation or Bayesian methods. Ankamah et al. [11] used VAR model to determine the impact of climatic variables on malaria. The analysis was conducted using monthly climate and malaria data from 2010 to 2015 in Ghana with the Granger causality test to examine the relationship between the two variables. VAR model was used with data mining method to extract knowledge, from January 2010 to December 2018. The variables obtained were tuition fee, inflation rate, number of enrolled students, and regional minimum wage [12]. Washington et al. [13] forecast the missing climate data using VAR model and the Root Mean Square Error (RMSE) for evaluation. Zhang et al. [14] also forecast the best oil production based on a multivariate time series (MTS) and VAR machine learning model for waterflooding. This method initially used MTS analysis to optimize injection and production data based on well pattern analysis. Subsequently, VAR model was established to mine the linear relationship from MTS data and forecast the oil well production by model fitting. Fitriani et al. [15] analyzed the development of COVID-19 cases and built VARI model issues in Indonesia and Singapore. Data used were from daily COVID-19 confirmed cases between March 16th and April 19th, 2020. Djara et al. [16] analyzed VARI model on data of Indonesia's exports and imports from January 2015 to March 2021, using Granger causality test to determine the relationship between variables. Sumertajaya et al. [17] also used missing data, which was applied to VAR model combined with Imputation Method (VAR-IM).

Based on previous investigations, this research aimed to VARI model for application in climate data, specifically case big data. VARI model was used to forecast climate in West Java Province with OLS as a parameter estimation tool. Climate parameter was rainfall and big data were obtained from NASA POWER. Subsequently, processing was carried out using the analytics lifecycle intended for big data problems and science projects.

II. Materials and Methods

A Data Analytics Lifecycle

Big data refers to extensive and complex datasets that are beyond the ability of traditional processing and analysis tools to manage and analyze effectively. Generally, data analytics lifecycle is explicitly designed for big data problems. It comprises discovery, data preparation, model planning, model building, communicating results, and operationalization. Data Analytics Lifecycle includes the recommended procedures for effectively managing the analytics process from initial exploration to project finalization. This lifecycle incorporates proven methods from the fields of data analytics and decision science [8].

Fig. 1. Flow of Data Analytics Lifecycle.

B Pearson Correlation

Correlation analysis is a statistical method used to identify a significant association between variables [18]. Pearson correlation is one of the tests used to determine the close relationship between two variables that have intervals or ratios, are normally distributed, and return the value of the correlation coefficient with a range of values between 0 and 1 [19]. The correlation value can be calculated using the following equation:

$$r = \frac{\frac{1}{n}\sum_{i=1}^{n}\left(X_i - \bar{X}\right)\left(Y_i - \bar{Y}\right)}{\left[\left(\frac{1}{n}\sum_{i=1}^{n}\left(X_i - \bar{X}\right)^2\right)\left(\frac{1}{n}\sum_{i=1}^{n}\left(Y_i - \bar{Y}\right)^2\right)\right]^{\frac{1}{2}}} \quad (1)$$

where

X_i : the value of X variable in time-t.

Y_i : the value of Y variable in time-t.

$\bar{X}$: average of X variable

$\bar{Y}$: average of Y variable

The Pearson correlation satisfies the following criterion that is shown in Table I.

TABLE I
PEARSON CORRELATION VALUE CRITERIA

r value	Classification
0.00-0.19	Very low
0.20-0.39	Low
0.40-0.59	Currently
0.60-0.79	Strong

C. Stationarities

The initial step in model identification in the analysis of time series data is testing the stationarity of data. The statistical test that determines objectively whether differences are necessary for stationarities of a time series is called the unit root test. When there is no significant change, data is considered to be stationary. Moreover, there are several kinds of unit root tests, among which the Augmented Dickey-Fuller (ADF) test is used as expressed below [10].

$$Z(t) = \phi Z(t-1) + U(t)$$

The steps used in the ADF test are as follows [20]:

1. Determine the statistical hypothesis test:
   $H_0 : \phi = 0$, data is not stationary in the mean (there is a unit root)
   $H_1 : \phi < 0$, data is stationary in the mean (no unit root)
2. Determine the error value (α).
3. Calculate the value of the test statistic using the OLS method:
   $$T = \frac{\hat{\phi} - 1}{S_{\hat{\phi}}} \sim t_{(n-p-1)}$$
4. Define test criteria:
   In the specified α there is a value $T > T_{tabel}$, then reject H_0, which means data is stationary or does not contain a unit root.

D. Differencing Process

According to previous research [10], the differencing process is a method of overcoming modeling on data that is not stationary concerning the average by differentiating or reducing the Z_t observation value with the value Z_{t-1}. For non-stationary time series data, the first-order differentiation process can be carried out using the following equation:

$$\Delta Z_t = Z_t - Z_{t-1} = (1-B) Z_t$$

where:
Δ : the first-order differentiation operator.
Z_t : observed value at time t.
Z_{t-1} : observed value at time $t-1$
B : Backward-Shift operator (Backshift)

When data is not stationary after applying order differentiation process, the next-order differentiation is carried out. Generally, the differencing process of consecutive order is carried out to obtain stationary data as follows:

$$\Delta^{\mathrm{d}} Z_t = (1-B)^d Z_t$$

E. Autocorrelation Function (ACF) and Partial Autocorrelation Function (PACF)

In the time series analysis method, the primary tool for identifying temporary model from data to be forecast is the Autocorrelation Function (ACF) and Partial Autocorrelation Function (PACF) through their correlograms. ACF shows the magnitude of the correlation or linear relationship between observations at the-t time (Z_t) and previous times $(Z_{t-1}, Z_{t-2}, \ldots, Z_{t-k})$. The equation for calculating ACF value is expressed below [21]:

$$\gamma_k = \hat{\rho}_k = \frac{\widehat{\gamma_k}}{\widehat{\gamma_k}} = \frac{\sum_{t=1}^{n-k} \left(Z_t - \bar{Z}\right)\left(Z_{t-k} - \bar{Z}\right)}{\sum_{t=1}^{n} \left(Z_{t-k} - \bar{Z}\right)^2}$$

PACF is the partial correlation between observations at the-t time (Z_t) and previous time $(Z_{t-1}, Z_{t-2}, \ldots, Z_{t-k})$. It can be calculated recursively using the following equation [21]:

$$\hat{P}_k = \frac{Cov\left[\left(Z_t - \widehat{Z_t}\right), \left(Z_{t+k} - \hat{Z}_{t+k}\right)\right]}{\sqrt{Var\left(Z_t - \hat{Z}_t\right)}\sqrt{Var\left(Z_{t+k} - \hat{Z}_{t+k}\right)}}$$

F. VARI)

VAR model is a multivariate time series model that can be used to examine objects with two or more variables influencing each other. It combines several AR model of the same order, forming a vector between VARIables that affect each other with strong correlation. VAR model was developed by Christopher A. Sims in 1980, with the general form of the order p of VAR(p) model stated as follows [21]:

$$Z_t = \sum_{k=1}^{p} \Phi_k Z_{t-k} + e_t$$

For VAR(p) which is not stationary, the differencing process is carried out k times until data is stationary. In this case, VAR(p) becomes VARI(p,k) which is expressed in the equation as follows [21]:

$$Y_t = \sum_{k=1}^{p} \Phi_k Y_{t-k} + e_t$$

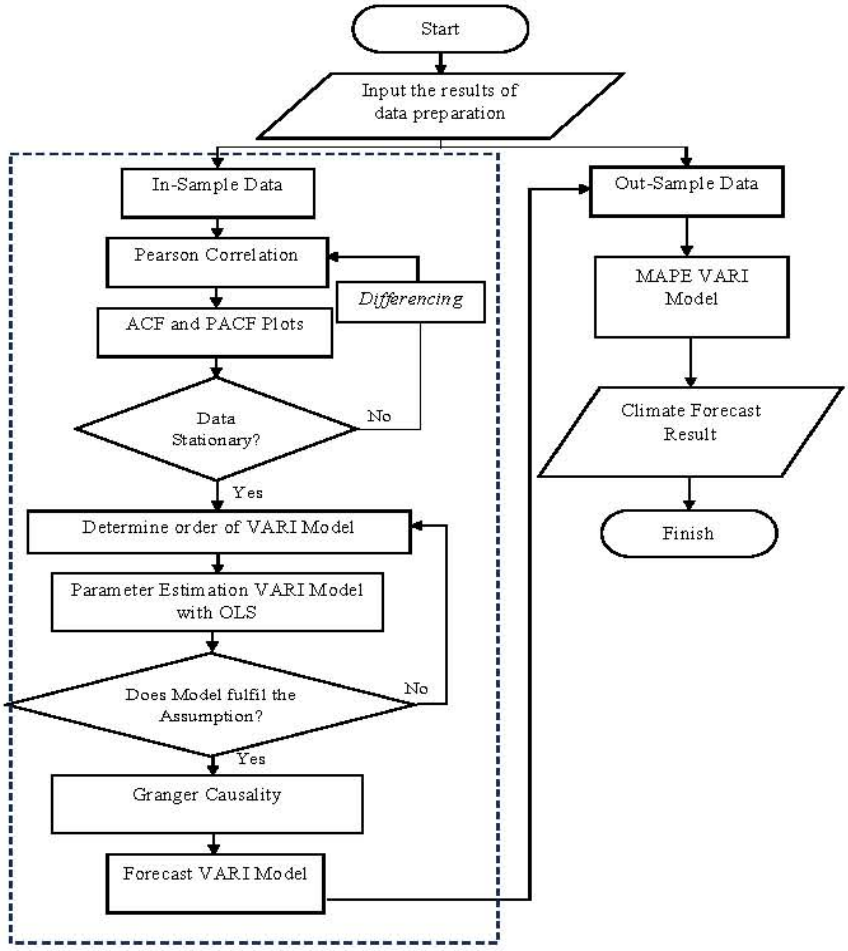


Fig. 2. Flow of VARI Model.

where $\mathbf{Y}_t = \mathbf{Z}_t - \mathbf{Z}_{t-1}, \ldots, \mathbf{Y}_{t-k} = \mathbf{Z}_{t-k} - \mathbf{Z}_{t-k-1}$.

VARI (1,1) model for N location is expressed in the following equation:

$$\mathbf{Y}_{t(N\times 1)} = \boldsymbol{\phi}_{p(N\times N)}\mathbf{Y}_{t-1(N\times 1)} + \mathbf{e}_{t(N\times 1)}.$$

In matrix form, it is stated as follows:

$$\begin{bmatrix} Y_{(1,t)} \\ Y_{(2,t)} \\ \vdots \\ Y_{(N,t)} \end{bmatrix} = \begin{bmatrix} \Phi_{11} & \Phi_{12} & \cdots & \Phi_{1N} \\ \Phi_{21} & \Phi_{22} & \cdots & \Phi_{2N} \\ \vdots & \vdots & \ddots & \vdots \\ \Phi_{N1} & \Phi_{N2} & \cdots & \Phi_{NN} \end{bmatrix} \begin{bmatrix} Y_{(1,t-1)} \\ Y_{(2,t-1)} \\ \vdots \\ Y_{(N,t-1)} \end{bmatrix} + \begin{bmatrix} e_{(1,t)} \\ e_{(2,t)} \\ \vdots \\ e_{(N,t)} \end{bmatrix}$$

Forecast process using VARI(1,1) model is shown in Figure 2.

G. Diagnostic Checking

Diagnostic checking is used to verify the assumptions that model must complete. These assumptions emphasize that there is no autocorrelation in the normally distributed residual model [10]. Model diagnostic tests can be performed using the Portmanteau and Jarque-Bera tests.

a. Portmanteau Test

The Portmanteau test is used to determine the autocorrelation of the residual model, as expressed in the equation below. The time series analysis assumes that the residuals must be independent (not correlated). When these assumptions are met, model is suitable for forecasting [10]:

$$Q = n\sum_{i=1}^{k} Y_k^2$$

The Q statistic follows the Chi-Square distribution with n^2k degrees of freedom. Therefore, the decision-making criteria are made by comparing the Q value with the Chi-Square value. When the value of Q > Chi-Square or p-value $< \alpha$, there is an autocorrelation in error.

b. Jarque-Bera Test

The Jarque-Bera test is used to determine the normality of the residuals, as expressed in the formula. When the residual normality assumption is met, model is suitable for forecasting [10]:

$$JB = \frac{n}{6}\left(S_k^2 + \frac{(k-3)^2}{4}\right)$$

The Jarque-Bera statistic follows the Chi-Square distribution with 2 degrees of freedom. Therefore, the decision criterion is made by comparing the Jarque-Bera

value with the Chi-Square value. When the value $JB >$ Chi-Square or p-value $< \alpha$, the residuals are not normally distributed.

H. Mean Absolute Percentage Error (MAPE)

MAPE is an evaluation tool of forecasting methods considering actual values' effect. As shown in Table II, a lower MAPE value correlates with higher accuracy of the method [22]:

$$MAPE = \frac{\left(\sum_{t=1}^{n} \frac{\left| Z_t - \overline{Z_t} \right|}{z_t} \right)}{n} \times 100\%$$

where

Z_t : actual value

$\overline{Z_t}$: forecast value

n : lots of observations

TABLE II
MAPE VALUE CRITERIA

MAPE	Accuracy
$MAPE \leq 10\%$	Very Accurate
$10\% < MAPE \leq 20\%$	Accurate
$20\% < MAPE \leq 50\%$	Reasonable
$MAPE > 50\%$	Not Accurate

III. Result and Discussion

Data analytics lifecycle method is applied to VAR model to forecast climate phenomena in West Java Province. The results obtained were as follows:

1. Discovery

Climate has an important role, as a significant change can affect human life. Therefore, climate phenomena are often investigated to provide an overview of climatic conditions. In this research, the phenomena explored were referred to as a time series, which allowed the use of VAR for prediction. Data analyzed were rainfall variables obtained from West Java Province. Based on classification, data were sourced from NASA POWER and included in big data with a distribution volume of 512,101.087 TB. Moreover, it was hypothesized that VARI model would provide an accurate forecast of climate phenomena based on the criteria for MAPE value of at least < 20%.

2. Data Preparation

Climate data collected from NASA POWER database were selected based on the observation location by inputting the latitude and longitude coordinates. Specifically, data used for analysis started from December 1989 until December 2022. The daily climate observations with the rainfall parameter was significantly large, resulting from 11,943 data for each location in 27 districts/cities. Furthermore, data uniformity and cleaning were carried out to handle missing values by removing noisy data containing errors such as NA values or non-numeric form data. After cleaning, transformation was performed by grouping based on aggregation. The daily data included in cleaning stage were aggregated into monthly to obtain 391 data for each location. This research collected data for December, January, and February (DJF), with 99 data for each location.

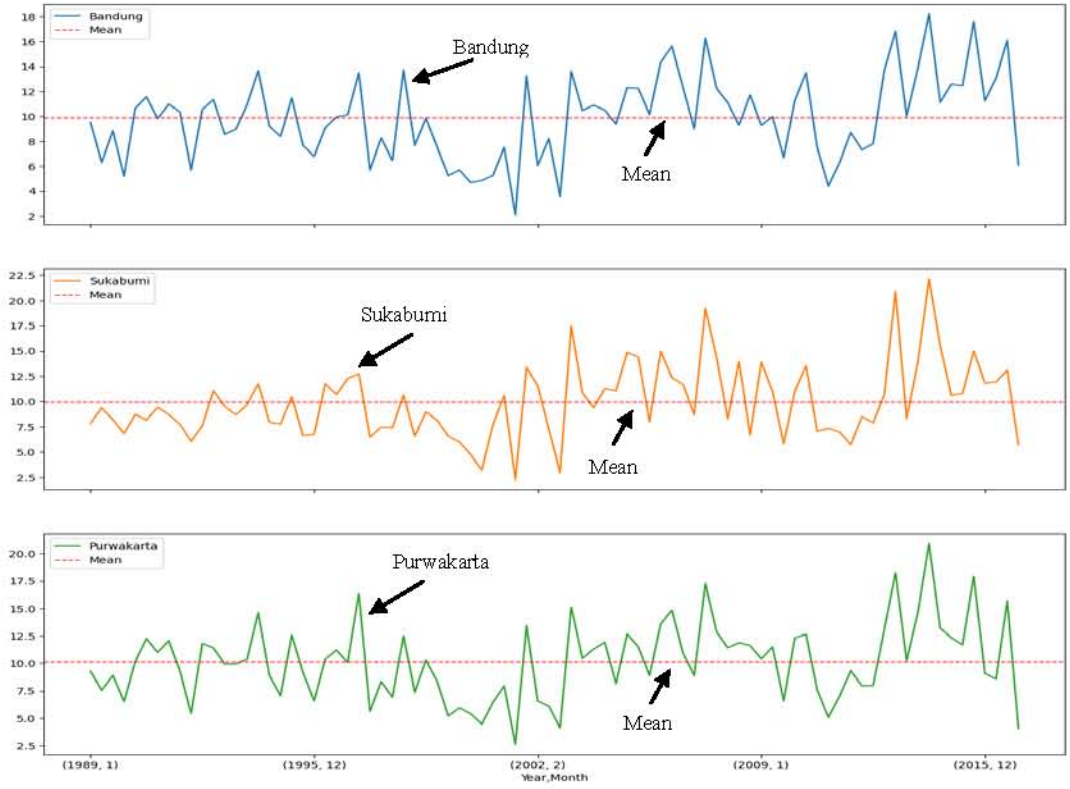


Fig. 3. Average Rainfall Data.

Subsequently, the selection of 3 locations representing the same observational values was carried out, namely the cities of Bandung, Sukabumi, and Purwakarta.

3. Model Planning

The results of data preparation are divided into in-sample and out-sample, comprising (89) 90% and (10) 10%, respectively. In-sample data were used to obtain VARI(1,1) model as shown in Figure 3, and out-sample was applied to forecast accuracy using MAPE value criteria.

In-sample data were input to determine the patterns on data plots. Based on observations, the red plot was Sukabumi data, the blue represented Bandung, and the green denoted Purwakarta, showing that data did not form a stationary pattern. This was followed by the calculation of Pearson correlation value to determine the relationship between locations, with the results presented in Table III.

TABLE III
PEARSON CORRELATION

	Bandung	Sukabumi	Purwakarta
Bandung	1	0.84	0.92
Sukabumi	0.84	1	0.84
Purwakarta	0.92	0.84	1

Based on Table III, the correlation between locations was included in the very strong category. Furthermore, the stationarity of data was checked using the ADF test, with the results shown in Table IV.

The order of VARI model was determined using the ACF and PACF plots in Figure 4. Based on the results, the lags are truncated at 1, 2, 3, and 11. In this case, the parsimony principle used produced order 1, resulting in the VARI(1,1).

TABLE IV
STATIONARITY

Location	*p*-value			
	Before Differencing	Exp.	After Differencing	Exp.
Bandung	0.089	NS	0.01	S
Sukabumi	0.074	NS	0.01	S
Purwakarta	0.198	NS	0.01	S

TABLE V
PARAMETER ESTIMATION

Location	Parameter	Estimator
Bandung	ϕ_{11}	-0.0140
	ϕ_{12}	-0.0238
	ϕ_{13}	0.0631
Sukabumi	ϕ_{21}	-0.0586
	ϕ_{22}	-0.0075
	ϕ_{23}	-0.0064
Purwakarta	ϕ_{31}	-0.1037
	ϕ_{32}	0.4952
	ϕ_{33}	-0.0026

4. Model Building

VARI(1,1) model estimation in this research used the Ordinary Least Square (OLS) method. The results of parameter estimation are shown in Table V and when substituted in VARI Mode,

$$Y_{(1,t)} = -0.0140Y_{(1,t-1)} - 0.0238Y_{(2,t-1)} + 0.0631Y_{(3,t-1)}$$
$$Y_{(2,t)} = -0.0586Y_{(1,t-1)} - 0.0075Y_{(2,t-1)} - 0.0064Y_{(3,t-1)}$$
$$Y_{(3,t)} = -0.1037Y_{(1,t-1)} + 0.4952Y_{(2,t-1)} - 0.0026Y_{(3,t-1)}$$

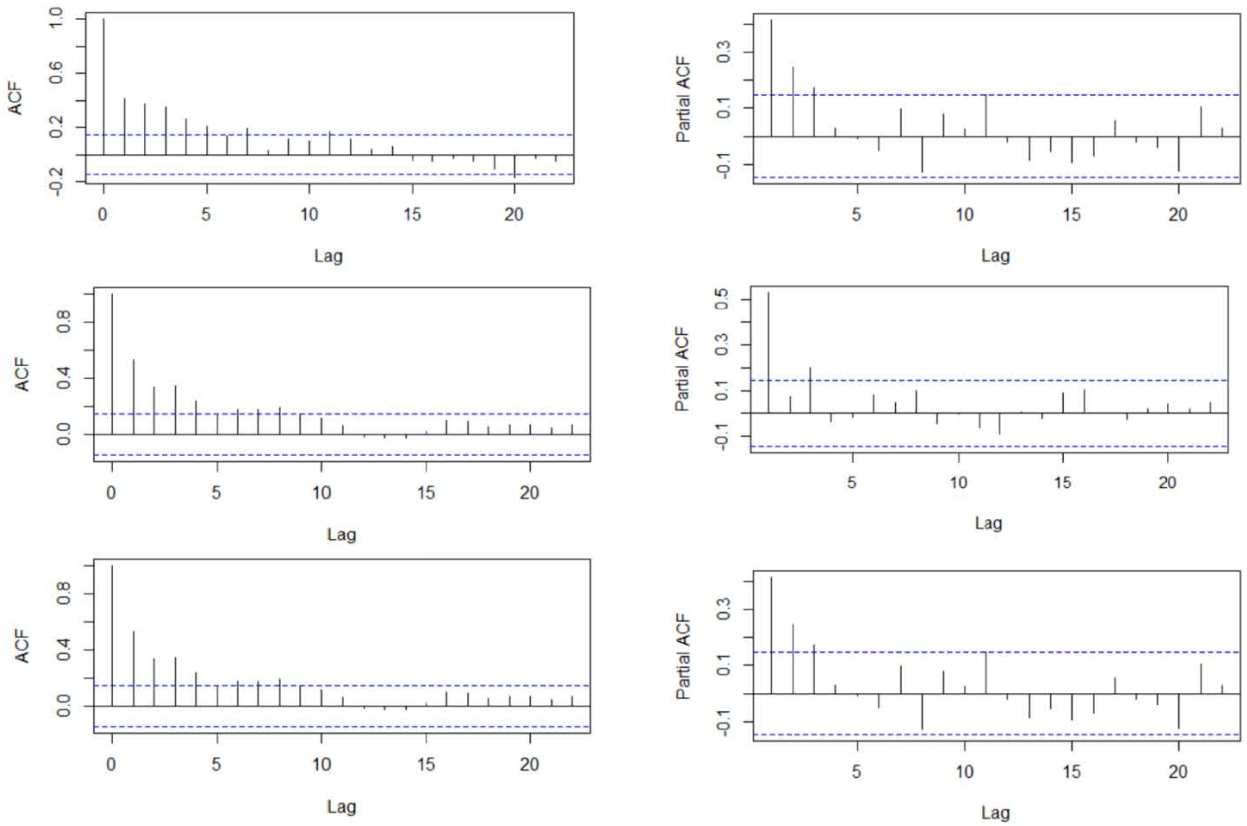


Fig. 4. ACF and PACF of Bandung, Sukabumi and Purwakarta.

where $\mathbf{Y}_t = \mathbf{Z}_t - \mathbf{Z}_{t-1}, \ldots, \mathbf{Y}_{t-k} = \mathbf{Z}_{t-k} - \mathbf{Z}_{t-k-1}$. Therefore, VARI model can be expressed in the following equation:

$$Z_{(1,t)} = 0.986Z_{(1,t-1)} + 0.0140Z_{(1,t-2)} - 0.0238Z_{(2,t-1)} + 0.0238Z_{(2,t-2)} + 0.0631Z_{(3,t-1)} - 0.0631Z_{(3,t-2)}$$

$$Z_{(2,t)} = -0.0586Z_{(1,t-1)} + 0.0586Z_{(1,t-2)} + 0.9925Z_{(2,t-1)} + 0.0075Z_{(2,t-2)} - 0.0064Z_{(3,t-1)} + 0.0064Z_{(3,t-2)}$$

$$Z_{(3,t)} = -0.1037Z_{(1,t-1)} + 0.1037Z_{(1,t-2)} + 0.4952Z_{(2,t-1)} + 0.4952Z_{(2,t-2)} + 0.9974Z_{(3,t-1)} + 0.0026Z_{(3,t-2)}$$

The Portmanteau test was carried out to determine the autocorrelation of model residuals, while Jarque-Bera was conducted to evaluate the normality. Based on Table VI, significance level was α=0.05, $p-value > \alpha$ in all test statistics. The results showed that data were normally distributed and no autocorrelation in the residuals, thereby, model could be considered feasible.

Granger causality check is used to check the causality relationship between locations. Significance level of $\alpha = 0.05$ showed that a causal relationship existed between locations, as presented in Table VII, with $p-value > \alpha$.

After fulfilling all assumptions, forecasting process was carried out using VARI(1,1) model for climatic phenomena with rainfall parameters by applying in-sample and out-sample data. The results presented in Figure 5 showed that actual and forecasting data plots had approximately the same pattern. Therefore, rainfall forecasting using VARI(1,1) model was considered a good method. The MAPE value also showed that the proportion of in-sample and out-sample data were 18% and 17%, respectively.

TABLE VI
DIAGNOSTIC CHECKING

Test Statistics	Accuracy	*p*-value
Portmanteu	66.8740	0.0931
Jarque-Bera	8.5750	0.0634
Kustosis	3.7420	0.0726
Skewness	6.9540	0.1450

5. Communicate Result

The results of calculated MAPE value for both in-sample and out-sample data was < 20%. Therefore, forecasting rainfall in West Java Province using VARI(1,1) model was included in the accurate category.

TABLE VII
GRANGER CAUSALITY

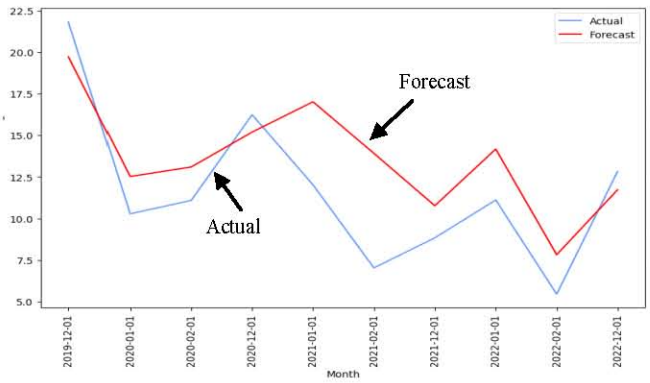
Location	*p*-value
Bandung	0.063
Sukabumi	0.065
Purwakarta	0.087

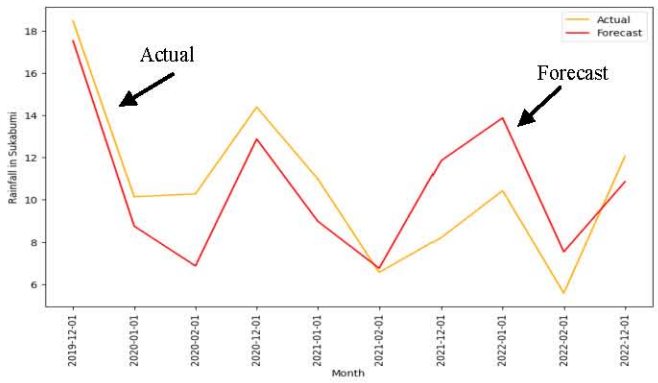
6. Operationalize

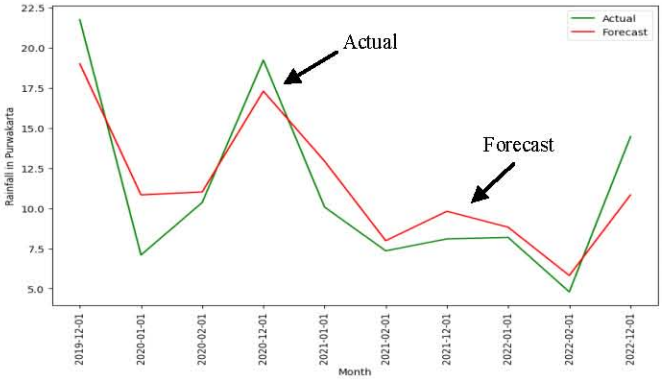
In practice, forecasting climate phenomena using the VAR model can be used as an early warning for local governments to minimize the impact of climate change.

IV. CONCLUSIONS

In conclusion, this research successfully investigated VARI model built with the assumption that data was not stationary. Model followed the Box-Jenkins procedure with identification using ACF and PACF, parameter estimation







applied to the OLS method, and diagnostic checking up to forecasting. VARI model was found to be a combination of the same order, which used order 1 in both AR and differencing process. Climate phenomena categorized as big data were processed using data analytics lifecycle method, which consisted of six phases, namely discovery, data preparation, model planning, model building, communicate results, and operationalize. The results showed MAPE value of less than 20%, indicating an accurate forecast.

ACKNOWLEDGMENT

Ajeng Berliana Salsabila is grateful to Prof. Budi Nurani Ruchjana and Prof. Atje Setiawan Abdullah for their valuable suggestions.

REFERENCES

[1] A. Amankwah, "Climate variability, agricultural technologies adoption, and productivity in rural Nigeria: a plot-level analysis," *Agric Food Secur*, vol. 12, no. 1, Dec. 2023, doi: 10.1186/s40066-023-00411-x.

[2] K. B. Dobbin, A. L. Fencl, G. Pierce, M. Beresford, S. Gonzalez, and W. Jepson, "Understanding perceived climate risks to household water supply and their implications for adaptation: evidence from California," *Clim Change*, vol. 176, no. 4, Apr. 2023, doi: 10.1007/s10584-023-03517-0.

[3] J. Huang, "Does Climate Risk Have the Same Impact on Corporate Bond Yields across Credit Ratings?" IAENG International Journal of Applied Mathematics, vol. 53, no. 2, pp. 566-572, 2023.

[4] W. W. Guo, L. Jin, W. Li, and W. T. Wang, "Assessing the vulnerability of grasslands in Gannan of China under the dual effects of climate change and human activities," *Ecol Indic*, vol. 148, Apr. 2023, doi: 10.1016/j.ecolind.2023.110100.

[5] BMKG, "Analisis Dinamika Atmosfer Dasarian II Mei 2023," May 23, 2023. https://www.bmkg.go.id/iklim/dinamika-atmosfir.bmkg (accessed May 28, 2023).

[6] WMO, "Climate," 2022. https://public.wmo.int/en/our-mandate/climate (accessed May 28, 2023).

[7] S. Sagiroglu and D. Sinanc, "Big data: A review," in *Proceedings of the 2013 International Conference on Collaboration Technologies and Systems, CTS 2013*, 2013, pp. 42–47. doi: 10.1109/CTS.2013.6567202.

[8] EMC Education Services, *Data Science & Big Data Analytics: Discovering, Analyzing, Visualizing and Presenting Data*. Indianapolis, Indiana: John Wiley & Sons, Inc., 2015.

[9] McKinsey Global Institute, "Big data: The next frontier for innovation, competition, and productivity," 2011. [Online]. Available: https://www.mckinsey.com/mgi.

[10] W. W. S. Wei, *Time Series Analysis: Univariate and Multivariate Methods*, Second Edition. Pearson, 2006.

[11] S. Ankamah, K. S. Nokoe, and W. A. Iddrisu, "Modelling Trends of Climatic Variability and Malaria in Ghana Using Vector Autoregression," *Malar Res Treat*, vol. 2018, 2018, doi: 10.1155/2018/6124321.

[12] S. Wahyuddin, F. I. Estiko, and E. Rijanto, "Analysis of Factors Affecting Tuition Fee in a Private University: A Data Mining Using VAR Model," in *IOP Conference Series: Materials Science and Engineering*, Institute of Physics Publishing, Nov. 2019. doi: 10.1088/1757-899X/662/2/022050.

[13] B. J. Washington and L. Seymour, "An adapted vector autoregressive expectation maximization imputation algorithm for climate data networks," *Wiley Interdiscip Rev Comput Stat*, 2020.

[14] R. ZHANG and H. JIA, "Production performance forecasting method based on multivariate time series and vector autoregressive machine learning model for waterflooding reservoirs," *Petroleum Exploration and Development*, vol. 48, no. 1, pp. 201–211, Feb. 2021, doi: 10.1016/S1876-3804(21)60016-2.

[15] R. Fitriani, W. D. Revildy, E. Marhamah, T. Toharudin, and B. N. Ruchjana, "The autoregressive integrated vector model approach for covid-19 data in Indonesia and Singapore," in *Journal of Physics: Conference Series*, IOP Publishing Ltd, Jan. 2021. doi: 10.1088/1742-6596/1722/1/012057.

[16] V. A. D. Djara *et al.*, "Prediction of export and import in Indonesia using vector autoregressive integrated (VARI)," *Journal of Mathematical and Computational Science*, 2022, doi: 10.28919/jmcs/7181.

[17] I.M, Sumertajaya, E. Rohaeti, A.H. Wigena, and K. Sadik, "Vector Autoregressive-Moving Average Imputation Algorithm for Handling Missing Data in Multivariate Time Series". IAENG International Journal of Computer Science, vol. 50, no. 2, pp. 727-735, 2023.

[18] U. Sekaran and R. Bougie, *Research Methods For Business: A Skill Building Approach*, Seventh Edition. John Wiley & Sons, 2016.

[19] Y. Zhang, Y. Li, J. Song, X. Chen, Y. Lu, and W. Wang, "Pearson correlation coefficient of current derivatives based pilot protection scheme for long-distance LCC-HVDC transmission lines," *International Journal of Electrical Power and Energy Systems*, vol. 116, Mar. 2020, doi: 10.1016/j.ijepes.2019.105526.

[20] A. Pal and P. K. S. Prakash, *Practical Time Series Analysis*. Packt Publishing Ltd, 2017.

[21] W. W. S. Wei, *Multivariate Time Series Analysis and Applications*, First Edition. John Wiley & Sons Ltd, 2019. [Online]. Available: http://www.wiley.com/go/wsps.

[22] K. D. Lawrence, R. K. Klimberg, and S. M. Lawrence, *Fundamentals of forecasting using Excel*. Industrial Press, 2009.